Vision Research 91 (2013) 62–77

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

What stands out in a scene? A study of human explicit saliency judgment

Ali Borji [a,*], Dicky N. Sihite [a], Laurent Itti [a,b,c]

[a] Department of Computer Science, University of Southern California, 3641 Watt Way, Los Angeles, CA 90089, USA
[b] Neuroscience Graduate Program, University of Southern California, 3641 Watt Way, Los Angeles, CA 90089, USA
[c] Department of Psychology, University of Southern California, 3641 Watt Way, Los Angeles, CA 90089, USA

ARTICLE INFO

Article history:
Received 13 August 2012
Received in revised form 9 July 2013
Available online 15 August 2013

Keywords:
Explicit saliency judgment
Space-based attention
Eye movements
Bottom-up saliency
Free viewing
Object-based attention

ABSTRACT

Eye tracking has become the *de facto* standard measure of visual attention in tasks that range from free viewing to complex daily activities. In particular, saliency models are often evaluated by their ability to predict human gaze patterns. However, fixations are not only influenced by bottom-up saliency (computed by the models), but also by many top-down factors. Thus, comparing bottom-up saliency maps to eye fixations is challenging and has required that one tries to minimize top-down influences, for example by focusing on early fixations on a stimulus. Here we propose two complementary procedures to evaluate visual saliency. We seek whether humans have explicit and conscious access to the saliency computations believed to contribute to guiding attention and eye movements. In the first experiment, 70 observers were asked to choose which object stands out the most based on its low-level features in 100 images each containing only two objects. Using several state-of-the-art bottom-up visual saliency models that measure local and global spatial image outliers, we show that maximum saliency inside the selected object is significantly higher than inside the non-selected object and the background. Thus spatial outliers are a predictor of human judgments. Performance of this predictor is boosted by including object size as an additional feature. In the second experiment, observers were asked to draw a polygon circumscribing the most salient object in cluttered scenes. For each of 120 images, we show that a map built from annotations of 70 observers explains eye fixations of another 20 observers freely viewing the images, significantly above chance (dataset by Bruce and Tsotsos (2009); shuffled AUC score 0.62 ± 0.07, chance 0.50, t-test $p < 0.05$). We conclude that fixations agree with saliency judgments, and classic bottom-up saliency models explain both. We further find that computational models specifically designed for fixation prediction slightly outperform models designed for salient object detection over both types of data (i.e., fixations and objects).

Published by Elsevier Ltd.

1. Introduction

Visual attention is a remarkable perceptual and cognitive capability of human visual system that selects and gates important information to higher-level cortical areas for further processing (see Baluch & Itti, 2011; Borji & Itti, 2013; Carrasco, 2011; Desimone & Duncan, 1995; Itti & Koch, 2001; Tatler et al., 2011, for reviews). Eye movements as proxies of visual attention have gained widespread use. They convey a lot of information about the processed scene regions when humans are engaged in free viewing tasks or performing daily activities such as sandwich making (Hayhoe, 2000). Previous research has shown two broad categories of visual attention mechanisms: exogenous *bottom-up* cues mainly based on characteristics of a visual stimulus (Nothdurft, 2005; Treisman & Gelade, 1980), and endogenous *top-down* cues determined by cognitive phenomena such as knowledge, expectations, reward, memory, goals, and task demands (Ballard, Hayhoe, & Pelz, 1995; Duncan, 1984; Land & Hayhoe, 2001; Navalpakkam & Itti, 2005; Navalpakkam et al., 2010; Posner, 1980; Yarbus, 1967). The relative influence of these two components varies across everyday behaviors. While for some behaviors attention is highly driven by task demands (Ballard, Hayhoe, & Pelz, 1995; Land & Hayhoe, 2001), for some others attention is believed to be influenced by bottom-up saliency (Itti, 2005; Itti & Koch, 2001; Parkhurst, Law, & Niebur, 2002; Peters et al., 2005; Tatler, Baddeley, & Gilchrist, 2005).

* Corresponding author. Address: University of Southern California, Hedco Neuroscience Building, 3641 Watt Way, Los Angeles, CA 90089, USA.
E-mail addresses: borji@usc.edu (A. Borji), sihite@usc.edu (D.N. Sihite), itti@pollux.usc.edu (L. Itti).

0042-6989/$ - see front matter Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.visres.2013.07.016

Eye movements, as one of the main ways to tap into visual attention, certainly convey a lot of information, but do not tell the whole story as attention is not necessarily always directed to the gaze location (i.e., covert attention, Posner, 1980; Wright & Ward, 2008). Yet, in using eye movements as a proxy for attention, it is often assumed that attention is primarily directed toward the location of gaze. For example, correlations between eye fixations during free viewing and saliency models have been the main criterion for judging how predictive models are (Foulsham & Underwood, 2008; Itti, 2005; Parkhurst, Law, & Niebur, 2002; Peters et al., 2005; Reinagel & Zador, 1999; Tatler, Baddeley, & Gilchrist, 2005; Zetzsche, 2005). Most models address the computation of salience and the guidance of covert attention without any consideration of gaze control mechanics. In addition, even in free viewing, one cannot rule out the existence of top-down factors. Thus, to justify using eye movements for evaluating saliency models, researchers have had to employ a variety of techniques. For example, rapid stimulus presentation times less than 5 s, focusing the analysis on the first few saccades that presumably are more bottom-up, and discounting center bias (Foulsham & Underwood, 2008; Parkhurst, Law, & Niebur, 2002; Tatler et al., 2011; Tseng et al., 2009) have been used to minimize the impact of top-down attentional components. While these techniques may help to some extent, they cannot completely eliminate conceptual top-down factors such as global scene context (Torralba et al., 2006) and semantic object dependencies (Hwang, Wang, & Pomplun, 2011) that influence the way people look at scenes. Top-down factors are even more abundant in free viewing of videos (Itti, 2005) where concepts such as actors, actions, predictions, social cues, gaze and pointing directions, and movement trajectories affect eye movements. Finally, free viewing fixations also confound slow and fast processing as they indiscriminately measure both truly saliency-driven saccades (presumably very rapid) and also slower memory/cognition-driven saccades (Powers, 2013).

From a physiological point of view, bottom-up and top-down attention have tight interplays. Fig. 1 illustrates the brain circuitry driving eye movements and shows that the superior colliculus, the last brain nucleus before the brainstem driving eye movements, receives inputs from a complex network, involving early and high-level cortical regions. This figure clearly shows that eye movements are driven by the joint of bottom-up and top-down attention. Experimental findings by Mannan, Kennard, and Husain (2009, 's) witness the presence of top-down influences in free-viewing. They studied the relative contribution of bottom-up and top-down influences and showed that fixations of normal observers conform less with predictions of the saliency map model (by Itti, Koch, & Niebur (1998)) compared with fixations of agnosia patients in free-viewing of natural scenes. These patients have severe problems recognizing objects and understanding global scene properties, thus having impaired ability in applying top-down guidance. Henderson and Hollingworth (1999) also argue that fixations during scene viewing tend to be idiosyncratic and influenced by both bottom-up and top-down factors. These and some other findings have triggered development of integrated attention models to better account for fixations. For instance target features have been used to bias attention during naturalistic search tasks (Ehinger et al., 2009; Navalpakkam & Itti, 2005; Zelinsky, 2008) or scene gist has been used to shrink the search space (Torralba et al., 2006). Ultimately when dealing with a black-box system like brain, where we do not have direct access to its entire neural mechanisms, a reasonable strategy is collecting more experimental (behavioral/neurophysiological) data. In the context of attention, we can study its bottom-up mechanisms by correlating all collected data with (imperfect) saliency measures that are known to be bottom-up by construction.

Above statements and findings hence imply that eye movements are contaminated by both top-down and bottom-up factors. This motivates us to consider other (complementary) possibilities of measuring saliency on complex natural scenes. We conduct an explicit saliency judgment task to explore whether human observers could access and report their internal sense of saliency directly, thus possibly discounting top-down factors through conscious introspection and decision. Note that we do not expect explicit saliency (or any single behavioral measure) to entirely replace other attentional proxies or eliminate all issues with testing attention, but we believe it can provide additional non-redundant information. We examine the validity of the hypothesis that humans are able to make saliency judgments (by checking observers' agreement) and

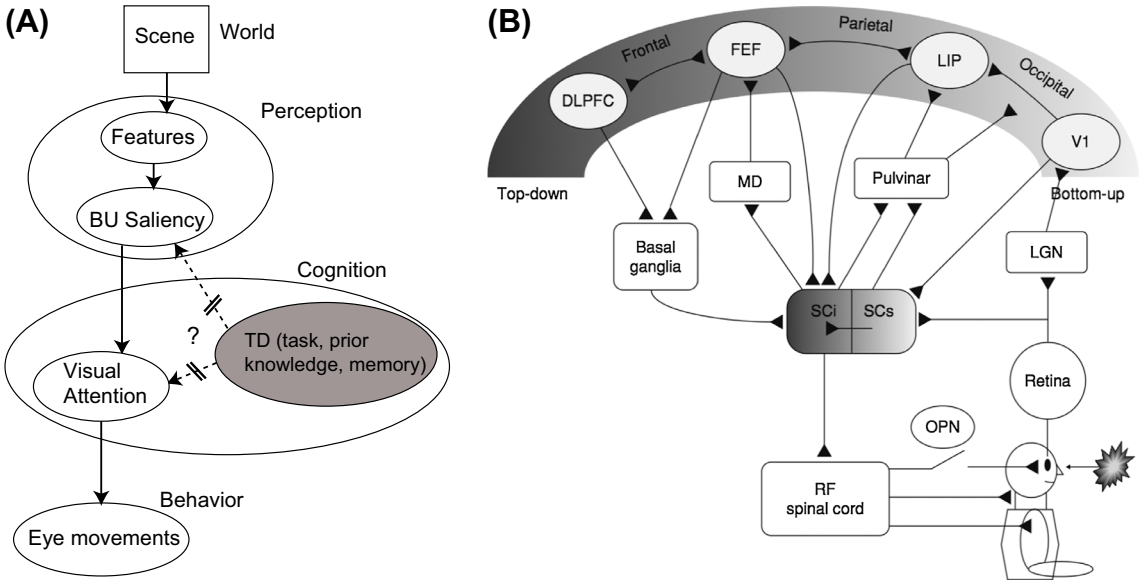

Fig. 1. (A) A simplified system overview of gaze control. Solid black arrows indicate the common assumption in free-viewing that eye movements are mainly driven by the bottom-up (BU) attention component. In daily life, top-down (TD) component of attention is always present. We ask whether observers can intentionally eliminate TD effects and suggest that, together with free-viewing, explicit saliency judgment can be used to study bottom-up attention. (B) Circuitry of the primate superior colliculus SC (adapted with permission from White and Munoz (2011)). Shading from light to dark represents the gradual shift from bottom-up to top-down processes respectively. SCs: SC superficial layers, SCi: SC intermediate layers. SC is influenced by both BU and TD components.

whether saliency models can explain such judgments (similar to free-viewing studies). While some behavioral results such as search accuracy or reaction time exist in pop-out visual search arrays, no study has thus far asked human observers for explicit judgment of conspicuous regions in natural scenes. Even those researchers addressing related concepts to saliency such as *interest* (Elazary & Itti, 2008; Masciocchi et al., 2009) mention that their task simultaneously reveals bottom-up and top-down influences on attention and they had no control over their relative contributions. In fact, Masciocchi et al. (2009) consider the presence of bottom-up and top-down factors in interest judgment as one reason why interest is a robust predictor of fixations in free-viewing (see additional details in Section 5).

In summary, we attempt to answer the following questions:

1. Are humans able to report saliency with little disturbance from top-down factors? Do they agree in their judgments? Do they have conscious access to lower-level fast saliency computations?
2. How can explicit judgments complement previous measures of visual attention? So far, saliency has been mainly measured through fixations, often with a free viewing task to minimize top-down influences, yet a major concern with this task is that there is no explicitly-asked and well-defined question for the observer (Henderson, 2003; Tatler et al., 2011). This could cause subjectivity (e.g., due to observer's mood, culture, interest, gender) and emanation of different top-down factors into eye movements. Our task has the advantage of being easy to conduct (no need for eye tracking, calibration, etc.).
3. Do saliency judgments correlate with fixations? If humans agree on their responses (hinting toward the objectivity of this task), one might expect explicit judgments and fixations, two indicators of attention, to be associated. Here we examine this expectation and analyze the degree to which explicit saliency correlates with fixations, by first employing all fixations and then each fixation separately (i.e., 1st fixations, 2nd fixations, etc.). While we do not expect a perfect correlation (due to differences in task, top-down set, etc.), some partial correlation would reinforce the idea that our explicit saliency measure can complement eye movement measures.
4. Can models of bottom-up attention explain saliency judgments? If so, which models correlate with saliency judgments better? Our experiments explore the generality of classic bottom-up saliency models in explaining other facets of attention in addition to free-viewing. Collecting new types of data also helps draw distinctions among existing models and investigate their biological plausibility.
5. Do explicit saliency judgments correlate more or less with object-based saliency models? Our tasks here probe the human saliency judgment at the object-level and not spatial locations. Using our data, we can compare models that detect and segment the most salient object in a scene versus traditional saliency models that predict fixation locations.
6. To what degree does object size correlate with saliency judgments? As a feature, *size* has been suggested to guide visual attention (e.g., Treisman & Gelade, 1980; Wolfe & Horowitz, 2004). How well can size alone predict which object might be selected as the most salient one, and how are predictions of saliency and size related?

As further detailed in Section 5, some studies have addressed related concepts that resemble our question at a first glance. There are, however, important differences: First and foremost, we explicitly address *saliency judgment*, not interest (Elazary & Itti, 2008; Masciocchi et al., 2009), object importance (Spain & Perona, 2010), nor memory recall (Einhäuser, Spain, & Perona, 2008; Isola et al., 2011). Please note that interest does not necessarily correspond to saliency as interest is a subjective concept while saliency is more objective which depends on low-level image features. Also, a salient object might not be interesting or important (and vice versa). Second, we address explicit saliency judgment at the object level (as opposed to Masciocchi et al. (2009)) as it is more likely that humans represent and understand a scene in terms of objects (Einhäuser, Spain, & Perona, 2009; Nuthman & Henderson, 2010). Therefore for explicit saliency judgment, it might be more natural for humans to choose objects (as opposed to clicking on salient points which needs some knowledge about center-surround dissimilarity). Note that interestingness has a large portion of top-down influences, while here we try to minimize them. Third, we carefully control the data gathering process and also use fixations (as opposed to Elazary & Itti (2008) who studied interest judgment using pre-collected data from the LabelMe dataset (Russell et al., 2008)). Above studies have used web-based tools to collect data where viewing distance, observer's engagement in the task, subject's posture or mood, and other factors are often hard to control.

From a computational perspective, our work reconciles salient object detection models with cognitive mechanisms of explicit saliency judgment in humans. We also conduct a model-based analysis by comparing both types of existing models (fixation prediction models such as Itti, Koch, and Niebur (1998) versus salient object detection models such as Goferman, Zelnik-Manor, & Tal (2010)) across two types of data. From an application perspective, in several occasions (e.g., image re-targeting or advertisement) it is more useful to know which objects explicitly stand out in a scene as opposed to predicting where people look.

2. Study of human explicit saliency judgment

We conducted two experiments in which observers were asked to "Select the object that stands out the most in the scene". We further explained to observers that they should select the object that maximally differs from the rest of the scene based only on its low-level features and visual appearance (and not conceptual top-down factors or preferences). The goal in the *first experiment* was to study the explicit saliency judgment of humans on simple stimuli which were images with only two objects. Observers had to click inside one of the two objects that they consider stands out the most.

In the *second experiment*, observers were supposed to draw a polygon around the object that stands out the most. We were concerned with the case of selection of a single salient object in an image. Observers' annotations were supposed not to be too loose (general) or too tight (specific) around the object. Observers were shown an illustrative example for this purpose.

2.1. Method

2.1.1. Participants

A total of 70 students (13 male, 57 female) from the University of Southern California (USC) participated in both experiments. The experimental methods were approved by the USC's Institutional Review Board (IRB). Observers had normal or corrected-to-normal vision and were compensated by course credits. Observers were in the age range between 18 and 23 (mean = 19.7, std = 1.4).

2.1.2. Stimuli and apparatus

For the *first experiment*, we chose 100 images of the SED dataset[1] by Alpert et al. (2007). Each image of this dataset contains

[1] This dataset is freely available online at: http://www.wisdom.weizmann.ac.il/~vision/Seg_Evaluation_DB.

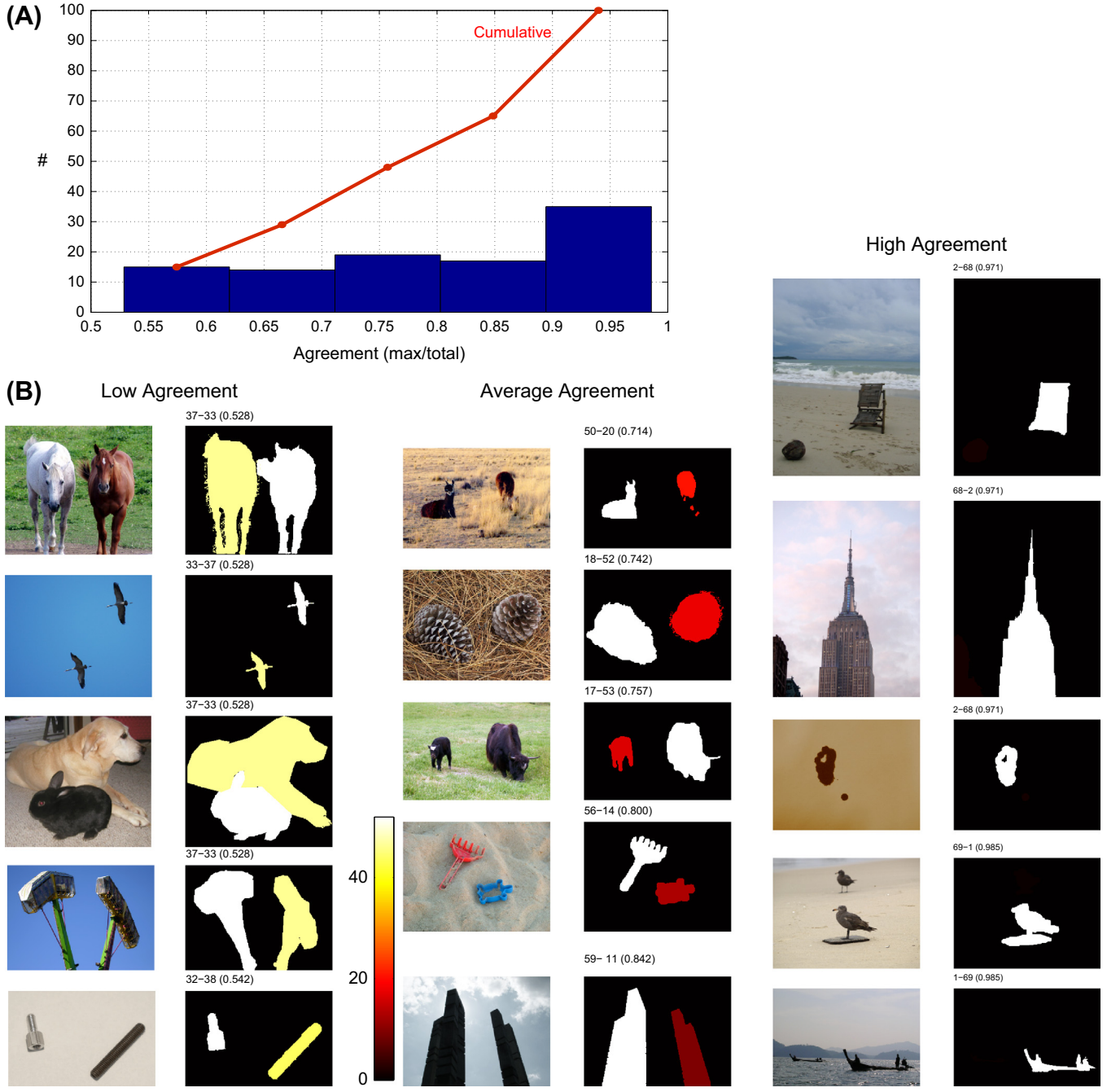

Fig. 2. (A) Histogram of saliency judgment agreement (max votes over total votes for each image; *r* values) according to Eq. (1). (B) Sample images with low, average, and high observer agreement along with their corresponding annotations.

two objects in a natural background with low clutter (e.g., grass or sky). Three human observers have already segmented object boundaries in these images. We then computed the union of these three segmentations and binarized it after thresholding (i.e., each pixel in the final map have been tagged by at least one of three observers).

For the *second experiment*, we chose the dataset by Bruce and Tsotsos (2009) which contains eye movements over 120 color photographs of indoor and outdoor environments with the resolution of 511 × 681 pixels. Images have been presented at random to 20 observers for 4 s with 2 s of delay (a gray mask) in between.[2] Observers in Bruce and Tsotsos's study viewed images freely.

Figs. 2 and 8 show sample images from the above-mentioned datasets.

In our experiments, observers were seated at a viewing distance of 130 cm from the screen (subtending a field of view of 43° × 25°). Stimuli were presented on a 42″ computer monitor at a resolution of 640 × 480 pixels and refresh rate of 60 Hz. These values were chosen to mimic the observer's distance of 0.75 m and screen size of 21″ used in the Bruce and Tsotsos dataset. We also attempted to match the color and luminance settings similar to Bruce and Tsotsos (2009).

2.1.3. Procedure

Both experiments were self-paced. In the *first experiment*, a trial ended when observers clicked inside one of the two objects. In the

[2] This dataset is freely available online at: www-sop.inria.fr/members/Neil.Bruce.

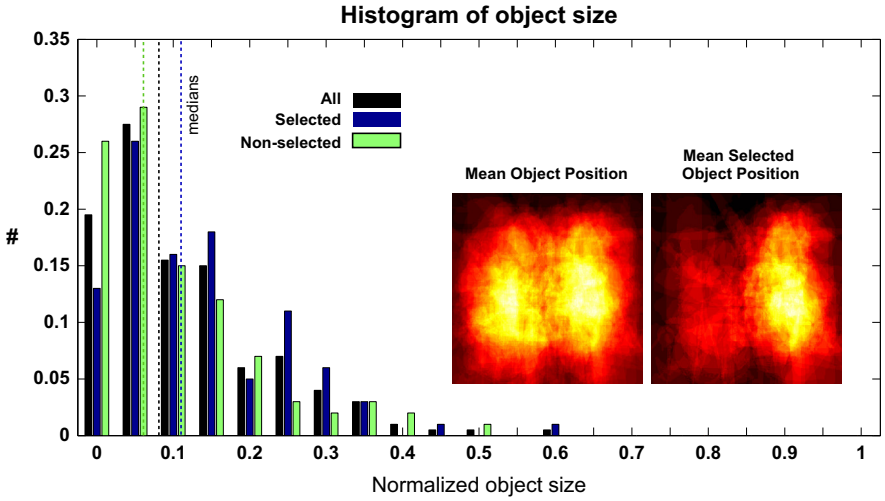

Fig. 3. Histogram of the normalized object sizes (object size over image size) for selected, non-selected and all objects. Medians of areas for selected, non-selected, all objects, and background (in order) are [0.107 0.061 0.083 0.818] (shown in dotted vertical lines). Median size of the selected objects is significantly higher than non-selected objects using *Wilcoxon ranksum test* which is the equivalent of *Mann–Whitney U-test* ($p = 0.006$). Mean object position for all objects and selected ones are shown in the inset. Salient objects judged by human observers are more frequent at the right side.

second experiment, observers successively clicked on the object boundary (to trace it) until they reached back to their starting point where a small circle hinted closure. They would then need to click inside the annotated object to move to the next screen. After the annotation, observers could get back to any point to further adjust their annotation (polygon). Observers had the opportunity to relocate their drawn polygon from one object to another. On average, each trial in the second experiment took 17.02 s (std = 3.66) over all images and observers.

Experiments were run back to back with no rest in between. Order of experiments were balanced across observers. The first experiment usually took about 10 min while the second one lasted about 35 min.

3. Results of Experiment 1

3.1. Quantifying observer agreement

To assess consistency in explicit saliency judgment, we define an agreement measure for observers' selection of the object that stands out the most. Let x_i and y_i be the number of votes (i.e., judgments) for two objects over all observers for the ith image. Then the agreement r_i is defined as $\max(x_i, y_i)/(x_i + y_i)$ with $x_i + y_i = 70$ being the number of all observers. Fig. 2A shows the histogram of r values and Fig. 2B shows images with low, medium, and high saliency judgment agreement. Observer agreement is above chance level (0.5) over all images. Object size is an important feature in judgments but subjects do not always base their decisions on this feature. In Fig. 2B, low agreement images often have two instances of essentially the same object (e.g., 2 birds, or 2 horses); if people had used an alternative strategy than saliency (e.g., prefer the larger object), then agreement should be high (e.g., see the large dog and small bunny image in the 5 images with lowest agreement). So, the low agreement in these images suggests that the two objects were closely tied in terms of their explicit saliency, and people were split close to 50/50 in picking one object or the other. For some images (Fig. 2B; right column), subjects tended to choose the larger object. We will analyze the role of object size in detail in Sections 3.4 and 3.6.

Fig. 3 shows the histogram of normalized object sizes (object size/image size) for *selected*, it non-selected, and it all objects. It shows a rightward shift for selected objects indicating that observers tended to choose larger objects. Fig. 3 also shows it Mean Annotation Position (MAP) defined as:

$$\mathrm{MAP} = \frac{1}{UV}\sum_{u=1}^{U=100}\sum_{v=1}^{V=70} s_{uv} \tag{1}$$

averaged over U images and V observers where s_{uv} is the annotation (i.e., vote) over the uth image by the vth observer. There are two peaks in the MAP map of images (at the left and right corresponding to two object positions; Fig. 3).

3.2. Quantifying visual saliency and employed saliency models

To quantify saliency, we exploit 10 state-of-the-art bottom-up saliency models that only use low-level image features (and not object detectors such as faces). The intuition behind using more than one model is to make sure that our results and conclusions are independent of the model type. Selected models include: *AIM* (Bruce & Tsotsos, 2009), *AWS* (Garcia-Diaz et al., 2012), *GBVS* (Harel, Koch, & Perona, 2006), *HouCVPR* (Hou & Zhang, 2007), *HouNIPS* (Hou & Zhang, 2008), *ITTI* (Itti & Koch, 2000), *ITTI98* (Itti, Koch, & Niebur, 1998), *PQFT* (Guo & Zhang, 2010), *SEO* (Seo & Milanfar, 2009), and *SUN* (Zhang et al., 2008). ITTI model is similar to the ITTI98 model but uses an iterative half-rectifying normalization operator which yields very sparse saliency maps. This variant of Itti et al.'s model is more desirable for machine vision applications as it clearly selects only a few salient regions in a scene. For more details on these models, the interested reader can refer to (Borji & Itti, 2013; Borji, Sihite, & Itti, 2013a). Note that *saliency* is not a unique measurement and may change from one model to another. That is why here we employ several models instead of one.

3.3. Comparing saliency of selected and non-selected objects, and the background

We hypothesize that bottom-up saliency inside the selected object is higher than the non-selected object. We evaluate the effect of *maximum saliency* inside the selected object on observers' decisions. Fig. 4A shows the hit percentage (using all models) for selected objects, non-selected objects, and image background for all

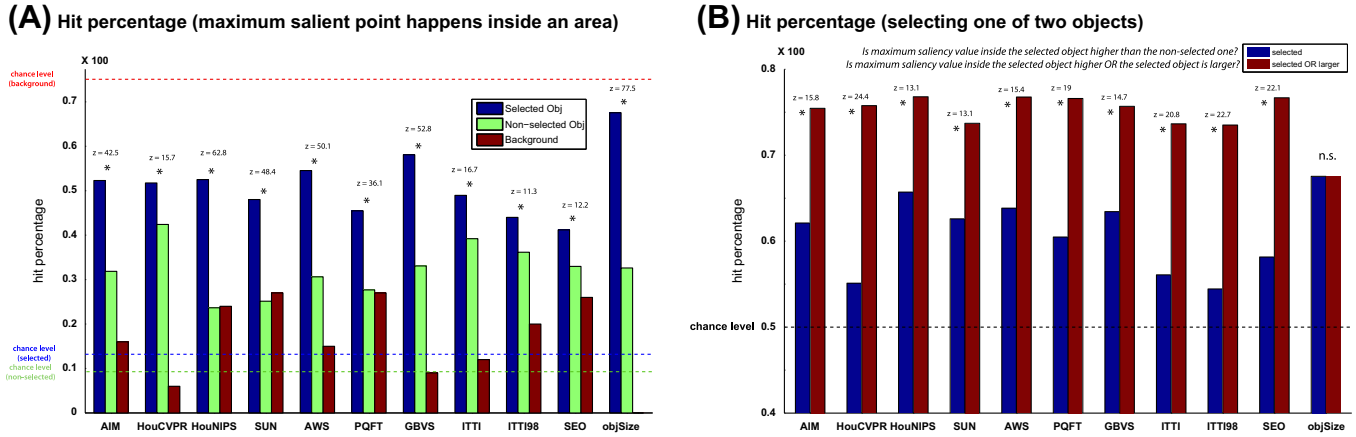

Fig. 4. (A) Hit percentage (i.e., the probability that the most salient point in the saliency map (of different models) happens inside the selected object, non-selected object, or the background). Saliency maps are linearly normalized to the [0 1] range. A star on each pair indicates significant difference using z-test ($\alpha = 0.05$). There is a significantly higher chance (for all models) for the maximum salient location to happen inside the selected object for each observer (separately). The same is true for comparison of the selected object versus background. For each model, hit percentages sum to 100%. Chance levels for all cases are also shown in the same color of their corresponding bars. Chance level is the probability that a point thrown randomly on the image to happen inside an area. This hit probability is proportional to the size of the target area. The *objSize* is the model that chooses the largest object. (B) Hit percentage by comparing the saliency of the two objects (versus human response). To test the statistical significance, the chance level for both decision criteria is a normal distribution (approximation of binomial) with mean 0.5 and standard deviation of 0.5 and thus the difference distribution has the mean 0 and standard deviation of 0.70. We then use z-test to compare the difference of two decision criteria (max saliency versus max saliency + size) with respect to this distribution. Differences are significant for all models. In another analysis to account for object size, we design a new decision criterion (red bars). A hit now occurs when either the selected object is the larger one or the most salient location happens inside it. Basically, if the most salient point always (or often) happens inside the larger object, then hit percentage of the combined rule should not increase (significantly), otherwise it means that both saliency and size are independent sources of information for observers' decisions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

observers and images. A hit happens when the maximum salient location happens inside an area. In some cases, upsampling the saliency map to the original image size, smoothness of the saliency map, or several equal maxima in the saliency map cause challenges for calculating a hit. To tackle this, we first find the maximum region. **If** the maximum region overlaps with the selected object, we consider it as a hit for the selected object, (**elseif**) if it overlaps with the non-selected object we consider it as a hit for non-selected object, (**else**) otherwise we count a hit for the background.

We check whether maximum saliency could be a predictor of human judgments. Let random variables X and Y represent events that a random point uniformly thrown in the image fall inside selected or non-selected objects, respectively. Each value is calculated for each image and each observer, separately. Each of these two variables has a binomial distribution which with many trials can be approximated by a normal distribution. We simulate a random process to measure mean and standard deviation of X and Y. We then define the random variable $Z = X - Y$ with $E(Z) = E(X) - E(Y)$ and $\sigma^2(Z) = \sigma^2(X) + \sigma^2(Y)$. Theoretically, means of the resultant normal distributions will be equal to the normalized object sizes.[3]

First, we test the hypothesis that maximum saliency is a predictor of human judgments and performs significantly higher than chance (i.e., $E(X)$). We calculate distributions of random variables by simulating a binomial process (i.e., by throwing a point uniformly random on the image and check whether it happens inside the selected object) 100 times for each image and observer pair, and calculating the mean and variance of the normal distributions (mentioned above). We repeat the same process for non-selected objects and for the background regions. For each model, by looping over images and observers, we calculate a 1×7000 vector (70 observers over 100 images; linearized) where each value indicates whether maximum saliency happens inside the selected object by that observer on that image, or not (similarly for non-selected objects and background). We then use the z-test to check whether our data samples (vectors) come from the same normal distribution represented by the random processes ($\alpha = 0.05$). Results are shown in Fig. 4A. We conclude that for all models, hit results (maximum saliency falling inside the selected object) are significantly higher the uniform random chance (same is true for non-selected objects). Maximum saliency does not fall on the background significantly higher than its corresponding chance level. This is in alignment with our expectation that salient regions happen on the objects rather than image background (Elazary & Itti, 2008).

Next, we investigate whether maximum saliency happens more often inside the selected object than the non-selected object. For the statistical test, we investigate whether the difference between hits on the selected and non-selected objects (looping over all images and observers) follows the distribution of variable Z or not (i.e., comparing with the difference in chance levels). We have a vector of size 7000 (selected - nonselected) with each element being either 1, 0, or -1 depending on whether the maximum salient point happens inside the selected object, background (none of the objects), or the non-selected object, respectively. Since variable Z has a normal distribution, we then test whether the above difference vector (for model hits) follows the distribution of Z or not using z-test. Shown in Fig. 4A, the difference is significant for all models ($\alpha = 0.05$). Thus, the hit percentage (averaged over observers and then images) of maximum saliency being inside the se-

[3] Variable Z has also a normal distribution since the difference of two independent (or correlated) normally distributed random variables has a normal distribution. We empirically and theoretically calculate parameters of the random variables. Over all images and observers, X has the mean of 0.1227 and standard deviation of 0.3281. Similarly, the mean and standard deviation for Y (i.e., non-selected objects) are 0.0721 ± 0.2586. These values for background regions are 0.8052 ± 0.4210. We also empirically calculate the $E(Z)$ by subtracting the vector Y from X, which is equal to 0.0506, identical to the theoretical mean (i.e., $0.1227 - 0.0721$). For variance, empirical variance ($0.3835^2 = 0.1471$) is smaller than the summation of variances (i.e., 0.1745). This means that the assumption that X and Y are independent is not true. However, we already know that sum of two normally correlated distributions is still normally distributed but the variance is no longer the simple addition of variances but is as follows: $\sigma^2(Z) = \sigma^2(X) + \sigma^2(Y) - 2\rho\sigma(X)\sigma(Y)$. Looking at the data, we empirically calculate the correlation coefficient ρ between X and Y which was 0.1617. Inserting this value in above equation, now both theoretical and empirical variances are exactly the same. There is a small correlation between sizes of the selected and non-selected objects (indeed in some images, both objects are two instances of a same kind, shown side by side; see for example horses and birds in Fig. 2).

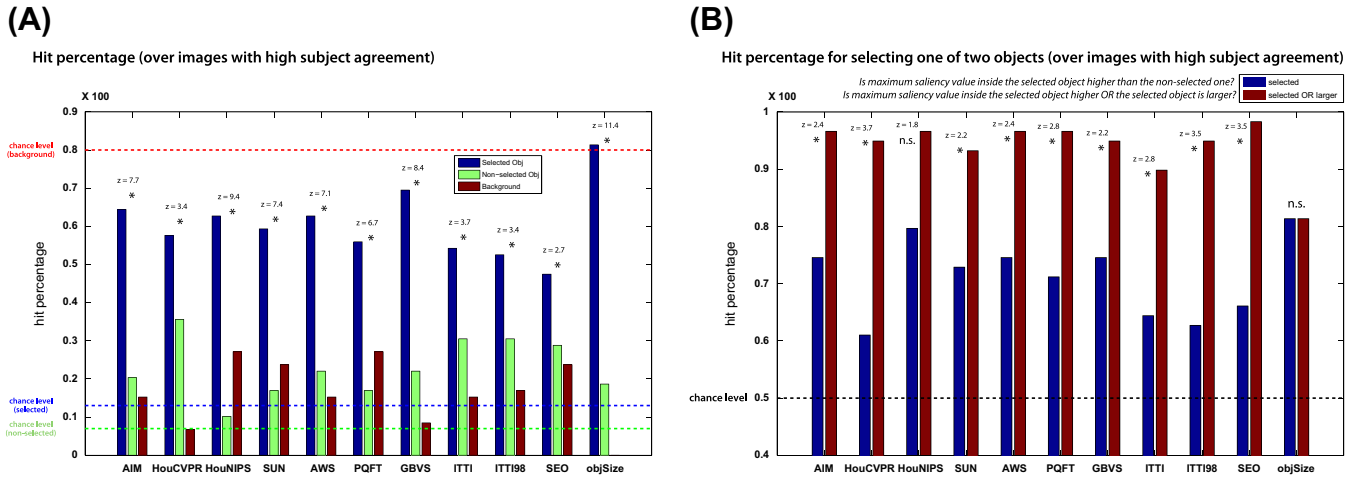

Fig. 5. Similar to Fig. 4 but over images with high observer agreement. (A) Hit percentage of maximum saliency for selected and non-selected objects, and background for images with observer agreement r greater than 0.75. Difference between selected and non-selected objects is significant for all models and is above results in Fig. 4. (B) Hit percentage of the *saliency combined with the larger object* rule enhances the performance of both factors significantly using z-test. Difference is only not significant for the HouNIPS model.

lected object is higher than the hit rate over non-selected objects (and also background) using all models.

To check observer variability, we perform a balanced one-way ANOVA (MATLAB) analysis. We calculate the accuracy of each observer (over all images, i.e., similar to Fig. 4 but for each observer separately). The result is a 70×3 matrix (denoted as A) where columns in order are accuracies for selected object, non-selected object, and background. Mean accuracies (across observers, each column) are significantly different ($df = 2$, $p < 5.76e{-}85$ for all models) which indicates that observers are consistent with each other (in accordance to Figs. 2A and 4A). To further investigate observer variability over selected versus non-selected objects (i.e., accuracy of saliency in predicting selections), we conduct a t-test ($\alpha < 0.05$) between the first two columns of matrix A (one matrix for each model). We observe a significant difference among selected versus non-selected objects across all observers ($p < 7.5e{-}14$ for all models).

3.4. Addressing object size

While the above analysis has already accounted for the confounding factor of object size, here, in a rudimentary analysis, we approach the problem from a different angle. From Fig. 3, we know that observers tend to select larger objects: the probability that the larger object in a scene to be selected is 0.1504 ± 0.1149, compared to 0.0851 ± 0.1149 for the smaller object being selected. The chance level that a random point falls inside the selected object (0.1286) is between these two values. To handle the possible confounding factor of object size, we devise a new decision criterion: a *hit* happens when the selected object is the larger one *or* the maximum saliency inside the selected object is higher than the non-selected one (i.e., contribution of both size and saliency together). With this new criterion (selecting one of the two objects), chance level[4] is now at 0.5. As Fig. 4B shows, hit percentages increase (compared with Fig. 3) meaning that both saliency and object size convey information regarding the selected object (again using z-test by comparing the difference between the new criterion and max saliency versus the difference of two chance levels which are normal distributions with mean 0.5 and standard deviation of 0.5). This is equivalent to focusing our analysis to cases where the selected object is not the larger one. Then maximum saliency can predict the selected object with the accuracy above chance (0.5).

Analysis of observer consistency for the combined rule of size and saliency (binary decision) shows a significant difference between saliency-only and combined rule predictions across observers (using first two columns of above-mentioned matrix A; t-test; $p < 8.4e{-}35$ for all models; corresponding to Fig. 4B).

3.5. Analysis of images with high observer agreement

In previous analyses, we used all images and observers but we did not differentiate between images with different levels of agreement (e.g., a 69 versus 1 vote against 36 versus 34). We considered a observer as an independent decision maker and tested a model's prediction against his decision. Here, we repeat the above analysis (i.e., Section 3.3) on images with high observer agreement (thresholding over the agreement r at 0.75 leads to 59 images; see Fig. 2). Now we have only one ground-truth answer for each image which is the selected object by the majority of observers (thus, there is no longer a loop through observers). A hit happens when the maximum value of a saliency map happens inside the selected object. We calculate parameters for random distributions and statistically test model hits (vectors of size 59) against them. Results are shown in Fig. 5A. Again, hit percentages by the maximum saliency are significantly higher than chance levels (probabilities are: selected = 0.1302 ± 0.03365; non-selected = 0.0698 ± 0.2549; background = 0.8000 ± 0.4000). Similar to above results, there is a significant higher chance for maximum saliency to happen inside the selected object than the non-selected object. The combined rule of maximum saliency and the larger object size outperforms both criteria significantly (Fig. 5B). Hit percentages are higher here compared with Fig. 4, indicating that when observers agree much, models also perform very well.

3.6. An integrated model of saliency and object size

So far we have shown that saliency and object size are important factors in observer's explicit judgment of the most salient object. How are these factors integrated by humans? Here, we suggest a model by linearly combining these two sources of information. We build a saliency map S which is equal to $\alpha S_{objSize} +$

[4] Please note that this is the accuracy of a completely random process that has no information about the object. Do not confuse this with the accuracy of the size model which we are considering as a baseline model.

$(1-\alpha)S_{model}$ where $S_{objSize}$ is the map with 1s at the location of the larger object in the image (annotation map), while S_{model} is the saliency map of a model. In practice, $S_{objSize}$ can be the output of a segmentation algorithm only for the largest segmented region. Parameter α determines the relative influence of two factors. We then vary α and calculate the hit percentage (i.e., fraction of the times that maximum of the combined map happens inside the selected object; by looping over images and observers). Results are shown in Fig. 6A for α ranging from 0 to 1 in steps of 0.02. In alignment with Fig. 4, results show that a linear combination of object size and saliency can explain the explicit saliency judgment of human observers. Fig. 4B shows accuracy of the combined model over all observers using three saliency models. Contribution of object size on observers' decisions varies across observers. Overall saliency influences judgments of majority of observers since there is at least one point where accuracy goes above the accuracy of the *objSize* model. Fig. 4C summarizes Fig. 4B by plotting the histogram of points where accuracy is maximum (i.e., α^*). While histograms differ across models, they show that saliency has been important in observers' decisions. Observers agree in the level of the contribution of size feature (in the combined model) for some saliency models (i.e., SUN or PQFT models show a large peak).

A sample image from the SED dataset, saliency judgments, and maps of 10 models are shown in Fig. 7.

4. Results of Experiment 2

Our purpose in the second experiment is to study human's explicit judgment of visual saliency over cluttered scenes containing multiple objects with different sizes. See Section 2 for experimental details. Fig. 8 shows sample images (smallest and largest) from the employed dataset (Bruce & Tsotsos, 2009) along with their annotations, histogram of normalized object sizes, as well as annotation agreement. Mean annotation map (over all images and observers) shows a center-bias indicating that in majority of images objects happen at the image center. The closest fitted bounding box to each annotated object has the average width (W) and height (H) of 95.78 and 100.74 pixels, respectively resulting in aspect ratio of 0.95 (W/H). This figure also indicates that there are few objects with large sizes in this dataset. Object sizes usually range from small to medium and occupy about 30% of the whole image.

4.1. Quantifying observer agreement

We define the following index to measure annotation agreement (or consistency in selecting object regions) among observers over the kth image:

$$r_k = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n=70} \frac{|s_{ik} \cap s_{jk}|}{|s_{ik} \cup s_{jk}|} \quad (2)$$

where s_{ik} and s_{jk} are annotations of ith and jth observers, respectively (out of n observers) over the kth image. Above index has the well-defined lower-bound of 0 when there is no overlap in segmentations of users and the upper-bound of 1 when they have perfect overlap. Fig. 8 shows histogram of r values. According to this figure, observers have low or medium agreement. Inspection of images with lowest agreement shows that these scenes have several salient objects while images with highest annotation agreement have often one unique salient object.

4.2. Correlation between explicit saliency and fixations

To test whether explicit saliency judgments (i.e., annotations) and fixations agree with each other, we treat each annotation map as a saliency map and use it to predict fixations. We use the Area Under the ROC Curve (AUC) (Green & Swets, 1966; Tatler, Baddeley, & Gilchrist, 2005) for evaluation of the fixation prediction power of an annotation map. For AUC calculation, first, the prediction map is resized to the image size where fixations have been recorded. Then, human fixations are considered as the positive set and some points from the image are sampled uniformly random as the negative set. The saliency map is then treated as a binary classifier to separate positive samples from negative samples. By thresholding over the saliency map, *true positive rate* is the proportion of fixations above a threshold while it false positive rate is the proportion of random points above that same threshold.

Fig. 9A shows the accuracy of the annotation map for fixation prediction. Annotation map scores AUC of 0.71 (std = 0.09) which is significantly above chance using t-test, $p < 0.05$. Chance level is the accuracy of a random map with value of each pixel drawn uniformly random between 0 and 1. We also report the accuracy of the Inter-observer (IO) model, a map build from fixations of other observers over the same image which is convolved with a small Gaussian kernel (see Fig. 11 for the size of the Gaussian).

4.3. Addressing the confounding factor of center-bias

One confounding factor when measuring the accuracy of a model against fixations is the center-bias. Center-bias is the tendency of human observers to preferentially look at the center of image (Parkhurst, Law, & Niebur, 2002; Tatler, 2007; Tseng et al., 2009). Due to center-bias in data, a simple Gaussian blob at the center of the image explains fixations better than almost all saliency models (Borji, Sihite, & Itti, 2013a). To tackle the center-bias issue, we use the shuffled AUC score (Tatler, 2007; Zhang et al., 2008) which is similar to the AUC score, but with the difference that negative points are randomly selected from fixations of other observers (plus the same observer viewing other images than the one under test) for each image (instead of being selected uniformly random over the entire image). Using the shuffled AUC, annotation map scores 0.62 (std = 0.07) which is again significantly above the shuffled AUC for the random map (t-test $p < 0.05$). Uniform random map scores 0.5 (std = 0.03) using both types of AUC scores. Overall, taking the results over two scores together, we conclude that human's explicit judgment of saliency and fixations agree with each other. In other words, people look at and judge the same object as the most salient object in a scene. Also, note that the prediction accuracy of the annotation map is as good as the ITTI98 saliency model.

4.4. Explicit saliency judgment and fixation order

We used all fixations in the previous analysis. Here, we attempt to know which fixations (i.e., first, second, etc.) match better with the annotation map. We repeat the same procedure as in Section 4.3, but each time evaluate the annotation map against a particular set of fixations on the image. Results are shown in Fig. 10 using shuffled AUC score based on the fixation order. Prediction accuracy is low at the first fixation, peaks at the 2nd one, and descends for subsequent fixations. It does not peak at the 1st fixation because majority of first fixations happen at the center of the screen (due to viewing strategy). This can be verified from average fixation maps shown in Fig. 10 (top of the figure). At the 2nd fixation (compared with the first one), observers have time to detect and direct gaze toward the most salient object. The accuracy of the IO model for all fixations is well above other predictors and shows a slight increasing trend. The AM map is as good as the AWS model and performs significantly above the ITTI98 model.

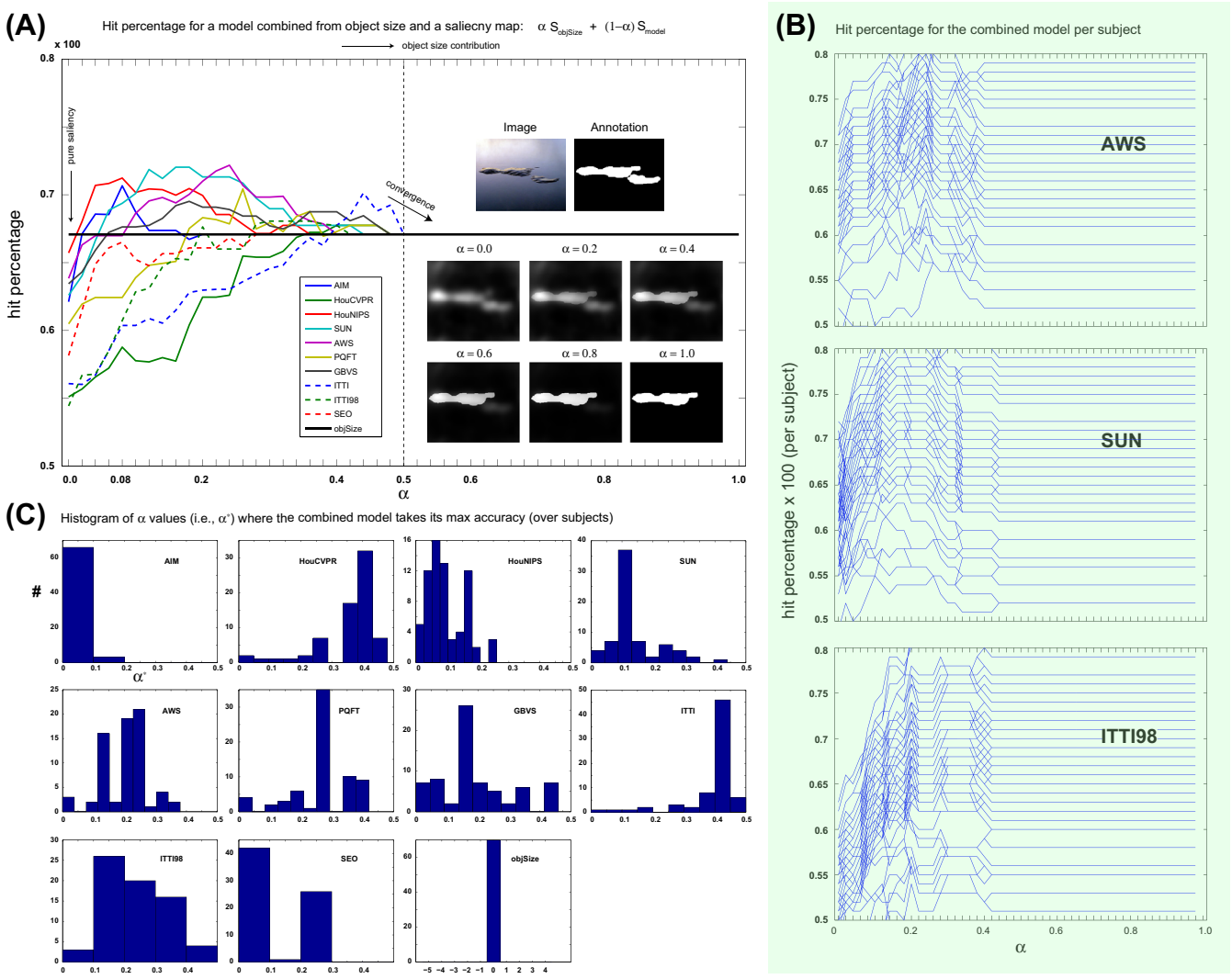

Fig. 6. (A) Hit percentages of our combined model of saliency (normalized to [0 1]) and object size for different levels of contributions of these two factors (α). For all models there is at least one point when accuracy surpasses the accuracy of the object size (objSize) model (the prediction of the larger object denoted by the horizontal solid black line). The maximum accuracy belongs to the AWS model at $\alpha = 0.24$. SUN model also results in high accuracy. For $\alpha > 0.5$, the object size model always wins the competition with the maximum salient location (because the term $1 - \alpha$ suppresses the saliency map) and thus performance flattens (converges) to the accuracy of the objSize model. Accuracies at the left ($\alpha = 0$) belong to the pure saliency (i.e., selection according to the maximum saliency criterion; see Fig. 4). Results are over all observers and images. A sample image, its annotation, and combined maps (larger object + AWS map) from some α levels are also shown. (B) Hit percentage of the combined linear model for each observer using AWS, SUN, and ITTI98 models. At one point in the x axis the accuracy of the combined model converges to the accuracy of the *objSize* model (i.e., selection of the largest object). These plots show that contribution of object size is different over observers. (C) Histogram of α^* values where the combined model (saliency + objSize) takes it maximum accuracy (over all 70 observers indicating a high observer agreement in employing object size).

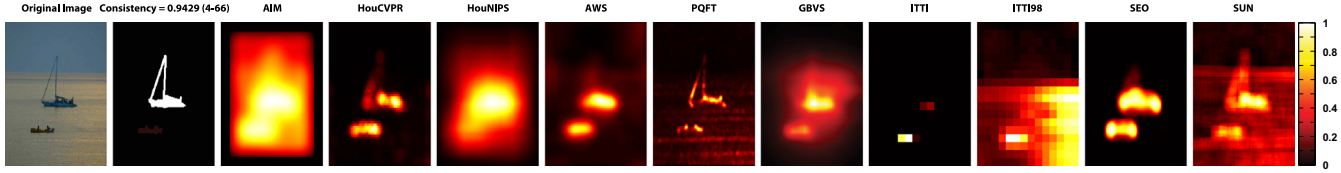

Fig. 7. A sample image, human consistency in saliency judgment, and prediction maps of employed saliency models.

4.5. Fixation prediction models versus salient object detection models

We conduct a model-based analysis by comparing the state-of-the-art models of salient object detection and fixation prediction over our data. On one hand, fixation prediction models need to accurately measure the conspicuity of local image regions in order to correctly predict humans fixation. On the other hand, salient object detection models try to generate homogeneous maps (similar to segmentation approaches) in a way that regions with high activation on these maps match with human annotation maps. We choose AIM, HouNIPS, AWS, GBVS, and ITTI98 models (used in Section 3) which have been shown to perform very well for fixation prediction in previous works (e.g., Borji, Sihite, & Itti, 2013a). For salient object detection, we choose Goferman (Goferman, Zelnik-Manor, & Tal, 2010), CBSal (Jiang et al., 2011), SVO (Chang et al., 2011), and RC (Cheng et al., 2011) models which have been shown

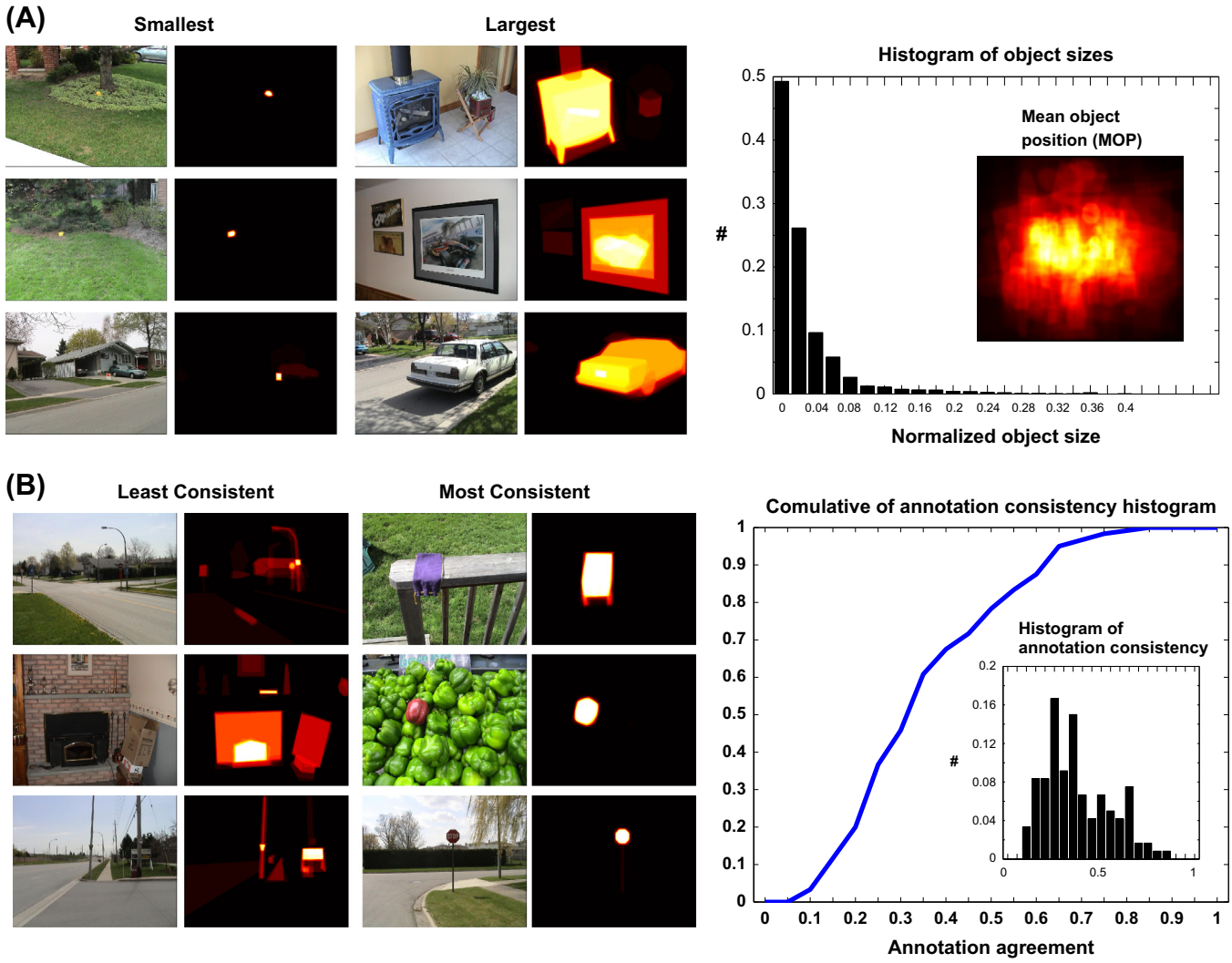

Fig. 8. (A) Left: Sample images with smallest and largest annotations. Right: Histogram of normalized annotated object sizes. More than 90% of the objects have sizes smaller than 10% of the image size. Mean annotated object map shows a strong center-bias (spatial prior). This is largely due to placement of objects at the image center (i.e., photographer bias; Tatler, 2007). (B) Left: Sample images with least and most annotation consistency. Right: Histogram of annotation consistency (i.e., r values; Eq. (2)). Consistency is higher for less cluttered scenes containing one or few salient objects.

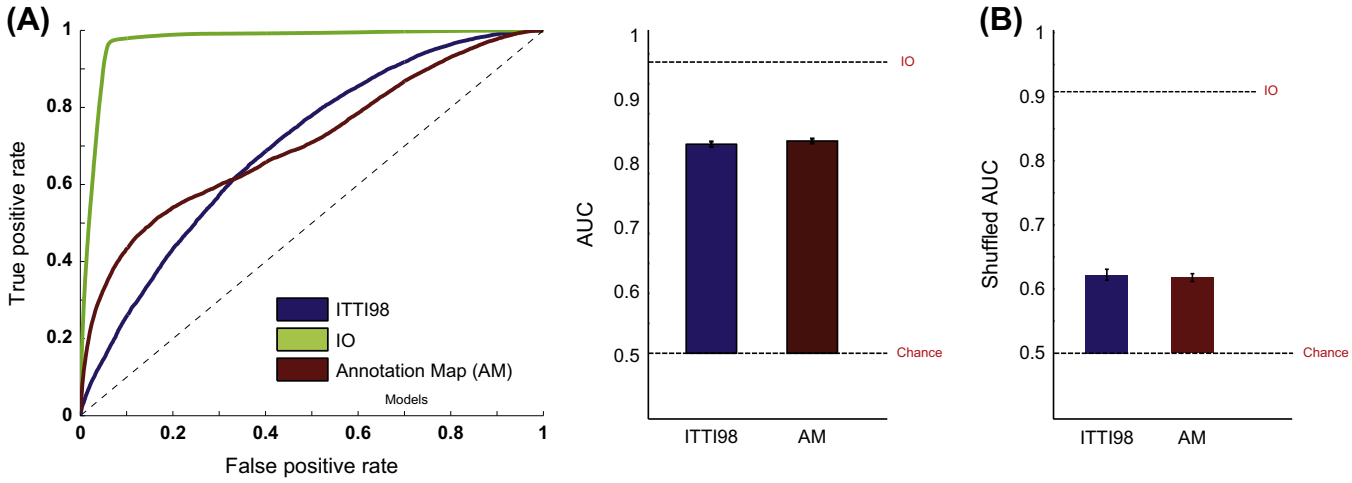

Fig. 9. (A) ROC curve and AUC values of the annotation map (see Fig. 8), Inter-observer (IO), and the ITTI98 saliency model for fixation prediction. Error bars indicate standard error of the mean (s.e.m) defined as: $\frac{\sigma}{\sqrt{m}}$, where σ is the standard deviation and $m = 120$ is the number of images (over images of Bruce and Tsotsos dataset). (B) Shuffled AUC scores of these models. The important point here is that the annotation map scores significantly above chance (i.e., AUC and Shuffled AUC of a random map are both equal to 0.5). AM model performs as well as the ITTI98 model. Note that the shuffled AUC values are smaller than AUC values due to discounting central bias in data.

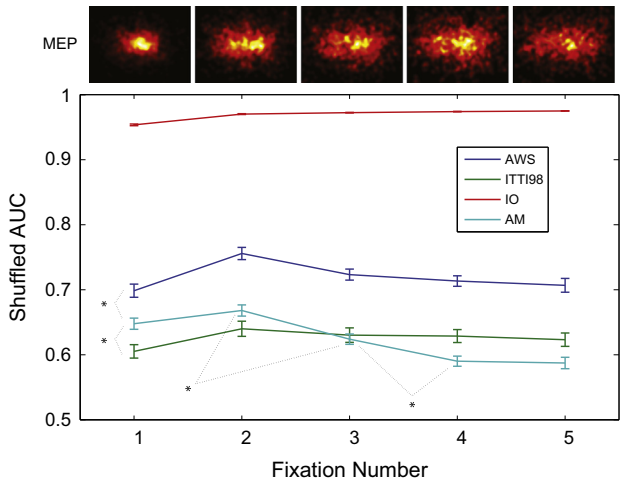

Fig. 10. Shuffled AUC of models over 1st, 2nd, 3rd, etc. fixations. Due to center-bias, scores are lower at first fixations. The difference between the AUC of the AM model at 2nd and 3rd fixations is significant (t-test; $p < 0.05$). Similarly for the 3rd and 4th fixations but not 4th and 5th. This means that observers tend to pick the most salient object at their second fixation. Top row shows the average fixation density for each fixation. AWS model stands on top among models. AM model scores significantly higher than the ITTI98 for first two fixations. MEP stands for the mean eye position map.

to perform very well (Borji, Sihite, & Itti, 2012). We apply both types of models to both types of data (fixations and annotations).

We use the shuffled AUC score to compare both types of models. To discount center-bias for salient object detection models, we simulate fixations over images similar to human fixations. This allows us to generate negative samples from the simulated fixations of other observers (and the same observer over other images) which is needed for calculation of the shuffled AUC score. For each image, simulated fixations are randomly drawn with a probability density equal to the annotation map (i.e., annotation map is considered as a probability distribution). This process is illustrated in Fig. 11 for a sample image. An advantage of the simulated fixations is that the same shuffled AUC algorithm used over human fixations is applicable here (which we actually use).

Results for both types of models are shown in Fig. 12A. Models originally built for fixation prediction (i.e., AIM, HouNIPS, AWS, and GBVS) have higher accuracy for predicting human fixations than salient object detection models (right bar chart in Fig. 12A). This is because fixation prediction models detect image-based outliers while salient object detection models try to merge salient regions into coherent object regions (i.e., segment the entire object and not its regions). Thus these two types of models somehow behave in the opposite direction. On the other hand, both categories of models have similar performances for prediction of simulated fixations (left bar chart). Over both types of data, there is a large gap between IO and models. This difference is larger over human fixations. In sum, our results show that despite the recent effort in computer vision for detecting the most salient object in a scene, still traditional fixation prediction models slightly outperform those models. This partially stems from the fact that current datasets used for evaluating salient object detection methods have images with often one salient object at the center, while here we employ cluttered scenes with multiple objects. Fig. 12B shows prediction maps for both types of models for a sample image.

Fig. 13 shows annotations and human fixations for sample images side by side. To overlay images with maps, we first normalize the map to the range of [0 1] and then multiply it with the image. For some cases, annotated objects fall at peaks of the fixation map while for some others there is not much overlap between two maps. Similar to Judd et al. (2009), we observe that humans consider faces, text, people, animals, and cars as the most salient objects in natural scenes.

5. Discussions

We now put our results into perspective with respect to the related literature mentioned earlier in the Introduction section.

5.1. Visual saliency, eye movements, and free-viewing

The traditional objective definition of saliency refers to bottom-up attentional processes that render certain image regions more conspicuous than the rest of the scene. Bottom-up saliency has been successful in predicting fixations in free-viewing of images (Foulsham & Underwood, 2008; Parkhurst, Law, & Niebur, 2002; Peters et al., 2005), videos (Itti, 2005), and visual search tasks (Ehinger et al., 2009; Treisman & Gelade, 1980; Wolfe & Horowitz, 2004). From a computational perspective, the concept of saliency has its roots in the early behavioral and psychophysical studies of visual search, Feature Integration Theory (Treisman & Gelade, 1980), and Koch and Ullman's computational architecture (Koch & Ullman, 1985). One of the early implementations of the saliency concept by Itti, Koch, and Niebur (1998) suggests that low-level feature discontinuities represented in the saliency map can explain a significant proportion of where people look. This is supported by studies that show measures such as local contrast correlate with fixation locations (e.g., Reinagel & Zador, 1999). In contrast to bottom-up saliency, top-down attention deals with high-level and cognitive factors that choose image regions relevant to a behavior, such as task demands, emotions, and expectations. For instance in visual search tasks, top-down attention biases search toward features of the target object (Ehinger et al., 2009; Navalpakkam & Itti, 2007; Rajashekar, Bovik, & Cormack, 2006). In more complex real-world tasks, such as sandwich making (Hayhoe, 2000) or driving (Land & Lee, 1994), where attention is tightly entangled with physical actions, fixations are driven to task-related locations to serve actions and goals. Please refer to Land and Hayhoe (2001), Tatler et al. (2011), Henderson et al. (2007, chap. 25), Navalpakkam and Itti (2005), Borji and Itti (2013), and Ballard, Hayhoe, and Pelz (1995) for a review of attention and eye movements in daily tasks.

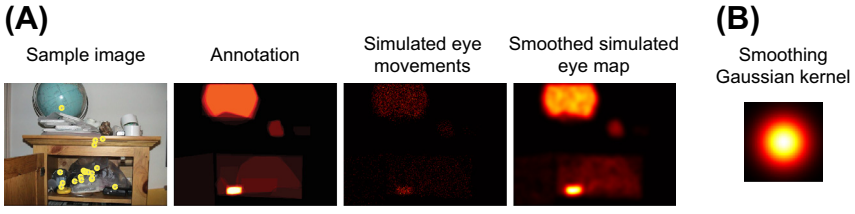

Fig. 11. (A) A sample image overlaid with human eye fixations, its annotation (averaged over all observers), simulated eye fixations (fixations are drawn with a probability proportional to the annotation map for that image; for each observer separately), smoothed version of the simulated map. (B) Gaussian smoothing kernel with size of 50×50 pixels (with $\sigma = 10$).

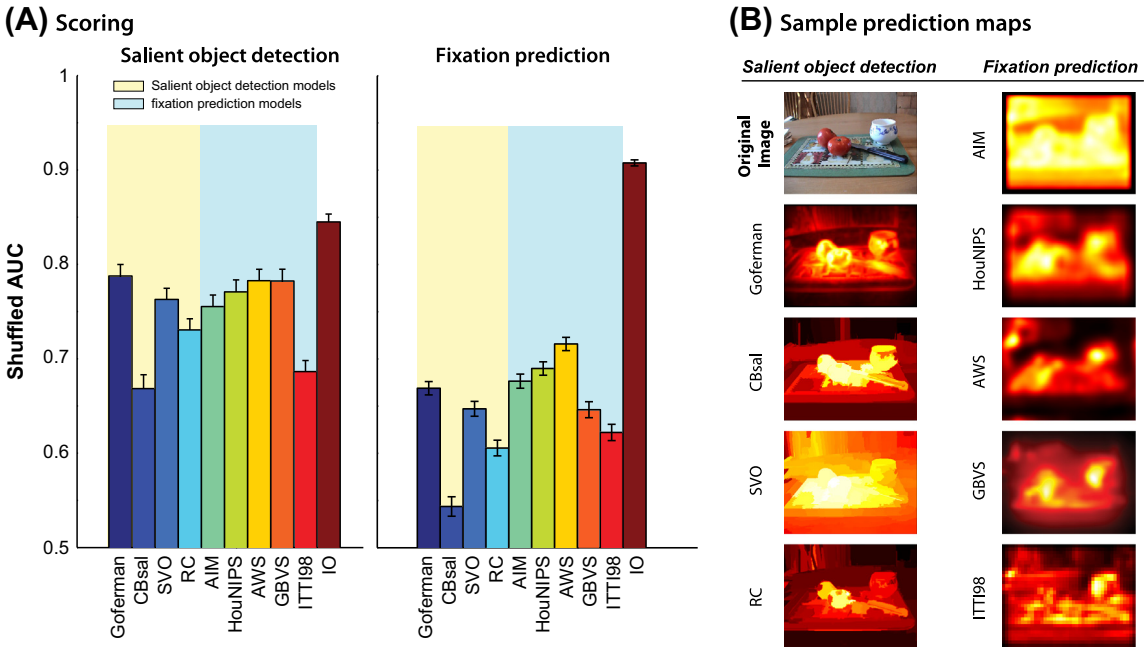

Fig. 12. (A) Fixation prediction accuracy of nine models (4 salient object detection and 5 fixation prediction) and the IO model (different IO models for human fixations and simulated eye movements). Shuffled AUC values for prediction of simulated eye movements (left bar) and human fixations (right bar) are also shown. Error bars are standard error of the mean (sem) over 120 images. (B) Prediction maps of models for a sample image.

Fig. 13. Sample images with their corresponding annotation and fixation maps overlaid.

Evaluating saliency hypothesis by measuring prediction power of models against fixations in necessary but not sufficient. Models can also be evaluated against some other findings about human attention. In this regard, our task offers such a basis. In addition, explicit judgment procedures can be used for study of covert top-down visual attention where humans attend to task-driven objects that they do not look at (e.g., walking in a sidewalk and avoiding mobile obstacles).

5.2. Objects and saliency

Our work is linked to previous object-based models of attention in that our definition of explicit saliency is inherently object-based. Subjects were asked to select salient objects based on whether the most salient location falls within their boundary. This suggests that salience map must have some object-level information incorporated. Some studies have emphasized the role of object information on guiding attention and fixations in scene viewing and other daily-life tasks such as grasping or manipulating an object, or to navigate an environment while avoiding obstacles. Einhäuser, Spain, and Perona (2009) investigated top-down and semantic selection of scene objects when humans freely viewed images for remembering their contents. They recorded eye positions while human observers inspected photographs of common natural scenes. Observers were asked to view an image and to name objects they remembered, right after the stimulus presentation. They showed that an annotated object map weighted by object recall frequency explains fixations better than chance. Hwang, Wang, and Pomplun (2011) showed that people tend to look at an object with less semantic distance to the currently attended one. As another support in favor of object-based representations, Nuthman and Henderson (2010) (and recently Pajak & Nuthmann, 2013) argued that if fixations are directed to objects, then they should happen around the center of objects. They were inspired by two findings: (1) people look at the center of words when reading text (Rayner, 1979) and (2) viewers prefer to look at objects over the background (Buswell, 1935; Elazary & Itti, 2008; Henderson, 2003; Torralba et al., 2006; Yarbus, 1967). Wischnewski, Belardinelli, and Schneider (2010) proposed a computational framework in which proto-objects, volatile medium-level static and dynamic features within the hierarchy of the visual system, are computed pre-attentively and are used to prioritize items of visual environment. The above findings are in line with object-based theories of visual attention stating that humans attend to objects and high-level concepts rather than spatial locations. Inspired by these cognitive results, some researchers (e.g., Cerf, Frady, & Koch, 2009;

Judd et al. (2009)) have used object detectors for objects such as faces, humans, or text to detect salient locations and predict fixations.

A separate line of evidence toward object-based attention comes from the visual search literature. Zelinsky (2008) proposed a model of visual search, known as target acquisition model (TAM) and was able to simulate human gaze patterns in search of toy objects using low-level object features. Rajashekar, Bovik, and Cormack (2006) showed that during visual search our attention and eye movements are biased by visual information resembling the target causing the image statistics near our fixated positions to be systematically influenced by basic visual features of the target. Mack and Eckstein (2011) proposed object co-occurrence as a contextual cue to guide and facilitate visual search in natural viewing. These studies along with Pomplun (2006) and Najemnik and Geisler (2005) suggest that object information have strong influences on the way we look at natural scenes. The first two results can however be also interpreted by those theories that postulate that attention modulates the weights of early visual features to render objects of interest more salient (Navalpakkam & Itti, 2007; Wolfe, 1998; Wolfe & Horowitz, 2004).

Eventually, the debate between two contrasting views: (1) attention is directed toward salient locations first and then a finer object processing is happening there, or (2) conversely saliency does not drive attention directly but through its correlation with objects, needs to be investigated by more controlled studies in future (see Borji, Sihite, & Itti, 2013a). One direction would be to eliminate spatial outliers and then see whether/how objects guide attention (i.e., replicating results of Nuthman & Henderson (2010) and Pajak & Nuthmann (2013) on textureless objects). It is however very likely that both spatial outliers and objects contribute to our allocation of attention and gaze in natural scenes. It is also likely that spatial image outliers direct attention heavily in absence of strong task demands. Alternatively, when there is a task, according to the cognitive relevance hypotheses (see Henderson, Malcolm, & Schandl, 2009; Tatler et al., 2011), objects may guide attention and fixations relatively more than spatial outliers.

5.3. Interest judgment

In an early study related to saliency judgment, Mackworth and Morandi (1967) recorded fixations of 20 observers over 2 images and asked 20 other observers for the recognizability (informativeness) of image patches (on a 10-point scale). Observers in the eye monitoring group fixated longer on image areas that observers in the rating group independently rated as more informative. Using a large database of labeled images, Elazary and Itti (2008) found low-level saliency a highly significant predictor of which objects humans choose to label (what authors considered *interesting* objects). The saliency map model found a labeled object 76% of the time within the first three predicted locations (chance = 43%). Masciocchi et al. (2009) addressed decision processes by which humans choose points in a scene as the most interesting ones (by mouse clicking on 5 most interesting locations). Using a large observer population (more than 1000 in a web-based online study), they found that interest selections are correlated with their eye movements, and both types of data correlate with bottom-up saliency. Masciocchi et al. (2009) concluded that interest and fixation judgments have both bottom-up and top-down attentional influences and that is the reason behind higher correlation between interest and free-viewing fixations.

5.4. Object importance

Spain and Perona (2010) addressed the problem of object importance and proposed a model for it. They argued that the goal of visual recognition is not only to detect and classify objects but also to associate a level of priority to each scene object. Since reliable algorithms do not exist for segmentation and recognition of all objects in a scene, Spain and Perona (2010, 's) approach has a limited applicability. Berg et al. (2012) proposed a less-constrained model of object importance which uses visual features and verbal descriptions on images. Spain and Perona (2010) also showed that bottom-up saliency influences object importance. Note that explicit judgment of saliency is not necessarily the same as object importance. Similar to Spain and Perona (2010, 's) study, we also verify that observers are able to make judgments at the object level. The impact of top-down factors seems to be more profound in importance judgment than saliency judgment, as importance judgment demands more semantic reasoning. Overall, all these three studies: Elazary and Itti (2008), Masciocchi et al. (2009) (by showing that clicks cluster around objects), and Spain and Perona (2010), support our finding that humans are able to judge saliency at the object-level. While importance, saliency, and interest are related concepts, further research is needed to elucidate their differences.

5.5. Salient object detection

Diverging from traditional saliency modeling for explaining fixations in free-viewing, Liu et al. (2007) attempted to detect and segment the most salient object in a scene. They asked 9 human observers to manually draw a bounding rectangle to specify a salient object in about 20,000 scenes (MSRA dataset). Achanta et al. (2009) annotated 1000 images from this dataset with mainly one unambiguous object (usually at the center) to set a benchmark for salient object detection approaches (ASD dataset). Although several models have been proposed for detecting salient objects in a scene (Borji, Sihite, & Itti, 2012), so far explicit judgment of saliency has not been systematically addressed in cognitive vision. Here we aimed to study this problem from both behavioral and modeling perspectives. Additionally, we share new data including annotations around salient objects chosen by human observers. As opposed to existing datasets, scenes in our dataset contain several objects in background clutter (on and off-center). Thus, our data offers a challenge for models, which have often been evaluated over images with a single object at the image center.

Thus far, several applications for salient object detection algorithms have been proposed in computer vision (e.g., image thumb-nailing (Marchesotti, Cifarelli, & Csurka, 2009), image compression (Itti, 2004), and object recognition (Kanan & Cottrell, 2010; Walther et al., 2005)). Behavioral investigation of explicit saliency judgment can further help practitioners in finding new application areas.

5.6. Mechanisms of explicit saliency judgment

Saliency is a pre-attentive and fast process allowing the brain to focus slower, complex, and expensive processes on few scene regions. We argued that humans have conscious access to a saliency map computed by the brain, since our observers were able to report reliably the most salient object in a scene. A likely mechanism is that humans have conscious access to the final saliency map, although maybe not to single feature maps (orientation, color, intensity, etc.) that contribute to saliency. An alternative possibility to conscious access to the saliency map is that observers keep track of where they look and finally choose an object among those they have visited in their scanpath. The latter suggests that judgments are made on the basis of a (re-) construction of conspicuity after effortful consideration of task requirements.

5.7. Size feature

Wolfe and Horowitz (2004) counted *size* as an undoubted attribute guiding attention. The size feature alone also explains a large fraction of our data (see Figs. 4–6). However, this feature is not fully predicted by models. Many saliency models have implicitly accounted for the size feature through incorporating multiple spatial scales of processing. But as we saw here, they failed to some extent, because they try to account for size over image regions and not over objects. Our results demonstrate that adding this feature increases accuracy of all employed models. Thus, models may be enhanced by addressing the size feature more explicitly. Further, explicit incorporation of Nuthman and Henderson (2010, 's) finding that humans tend to look at the center of objects may help increase accuracy of existing fixation prediction models.

5.8. Advantages and disadvantages of explicit and implicit saliency judgments

Explicit saliency judgments can supplement other previous measures (e.g., eye movements, reaction times, accuracy) for two main reasons: *First*, some researchers question the very nature of free-viewing as there is no explicit task nor a well-defined question[5] and it is almost impossible to cut all top-down influences. This could potentially cause idiosyncrasies in observers' behaviors as top-down influences might be very subjective and dependent on such factors as time of the day, mood, cultural preferences, language, and gender (Shen & Itti, 2012). As a result, it is not enough to use fixations in free-viewing tasks for probing bottom-up saliency. Here, we directly and systematically tackled this confound by asking observers to explicitly determine the most salient object in a scene. We further asked them to base their answers on low-level image features irrespective of high-level factors such as importance, interest, and context. *Second,* while the automatic and rapid nature of eye movements in free viewing initially made them attractive to study attention (Parkhurst, Law, & Niebur, 2002), recent research has challenged this idea. Indeed, a number of pitfalls of eye movement studies have recently been identified, such as the influence of context, center bias (by which humans tend to preferentially look towards the center of an image), and others (see Borji, Sihite, & Itti, 2013a; Parkhurst, Law, & Niebur, 2002; Tatler, 2007; Tatler et al., 2011). Although our explicit tasks do not guarantee to eliminate biases, the longer stimulus presentation time and the self-paced nature of our tasks here (compared to free-viewing) allow observers to fully inspect the image and to choose the object they think is the most salient one, possibly allowing them to voluntarily reduce such biases. Our task can also help dissociate between covert and overt attention (Posner, 1980; Wright & Ward, 2008). When someone does a visually-guided task, sometimes s/he might not look at something but still pay attention to it. In such case, eye movement recordings cannot reveal which object was attended. But, when asked, subjects may be able to report it explicitly.

[5] For example Tatler et al. (2011), Henderson (2003), and Henderson et al. (2007) believe free viewing is simply not a representative behavior and has limited applicability to what we really do with our visual system. Some other studies (e.g., Land & Hayhoe, 2001; Tatler et al., 2011; Triesch et al., 2003) argue that saliency is not a strong predictor of fixations when there is a task. In Borji, Sihite, and Itti (2011), we have objectively confirmed this argument. While we believe that top-down factors are important and always present, this does not necessarily mean that free-viewing never happens is real life. A more realistic viewpoint is that top-down factors are always present but their contribution is dependent on task demands and varies over time as task proceeds. High task demands (e.g., negotiating a turn while driving) may more strongly determine where eyes should be guided. There are occasions when task demands are so low that observers may explore the scene more freely. As claimed here and from the previous literature, bottom-up saliency more strongly drives fixations and attention in such cases.

The high degree of consistency among humans in selection of the most salient object in our experiments suggests that humans have an objective criterion, rather than idiosyncratic decisions for determining salient objects or regions. This means that our explicit saliency task is a good complement to free viewing and to eye movements for studying attention. This is good news as it now enables new experiments, such as large-scale web-based studies, to be run. In addition, laboratory setups will be easier and there will be no need for an eye tracker, calibration, etc. Another advantage of our proposed task is that it eliminates several challenging factors for example, uncertainty in eye-tracker accuracy, uncertainty of making a saccade by the observer (saccade end-point), smoothness of fixation maps, scoring, etc. Asking humans for their judgment at the object-level may be even more natural than asking them to click on salient or interesting points (Masciocchi et al., 2009). One may indeed argue that humans assign interestingness to objects rather than spatial points in the image. Besides, judging salient points needs a finer knowledge of spatial outliers or center-surround mechanisms which are implemented in saliency models but are difficult for humans. Finally, for some applications it is more advantageous to know which object people explicitly find is most salient (e.g., advertisement design) as opposed to where they look.

While explicit judgment tasks offers insights to the nature of human attention and model comparison, similar to eye movements, they have some limitations. They do not completely eliminate subjective idiosyncrasies as different subjects might have different takes on low-level features when asked to choose salient objects. In some other scenarios, however, this might be less of a problem. For example, when asked to explicitly judge the most task-relevant object that should be attended, subjects perhaps agree more with each other. This is partly because the task demand and the objective function is stronger compared to free-viewing of natural scenes. Another problem with our explicit judgment task is subject's accuracy in annotating object boundary. Although subjects were shown a representative sample, some subjects still may have chosen tighter or looser boundaries. Some subjects may have chosen to annotate smaller objects or easier ones.

6. Conclusions

In our first experiment, classic bottom-up saliency and size features explain about 80% of human judgment (Fig. 4). This suggests that humans were good at understanding the task instruction. In the second experiment, the shuffled AUC score of 0.62, prediction accuracy of the annotation map for explaining eye fixations, means that majority of fixations happened on the most salient object. We find that humans tend to find and segment the whole extent of the object while eyes are mainly driven to image outliers which may not necessarily form objects. We also find that human annotations often fall on conceptual items which are also salient, usually containing a full object or a part of it (e.g., faces). This finding matches with behavioral findings from eye movements studies (e.g., Cerf, Frady, & Koch, 2009; Judd et al., 2009) that faces, text, people, etc. attract human fixations. Some of these objects have been argued to be bottom-up salient due to their features (e.g., text and maybe faces). Faces are specially interesting because of their evolutionary importance witnessed by the fact that certain brain areas are devoted to process them (face cells in fusiform gyrus and IT cortex) as well as facial expressions and emotions. For other objects such as people, cars, and animals the landscape is not much clear and selection might be because of their low-level features.

In summary, we studied the saliency judgment of human observers when they were explicitly asked to choose the most salient object in a complex scene. Our investigation in experiment one

revealed that humans agree in their judgments and have conscious access to bottom-up saliency (Fig. 2). We further showed that classical bottom-up saliency models, detecting spatial image outliers, correlate with humans' judgments of the most salient object in a scene in terms of low-level features. In experiment two, we showed a high correlation between explicit saliency and fixations while humans freely viewed natural scenes. As opposed to previous studies which have usually used one saliency model (often model by Itti, Koch, and Niebur (1998)), here we employed several state-of-the-art models to validate independence of our results on model type. We discounted the center-bias factor to make sure that correlation between fixations and explicit judgments (annotations) is not because of high distribution of objects or fixations at the image center (a phenomenon known as photographer bias; (Tatler, 2007)). We categorized the existing models of bottom-up saliency as those that resemble segmentation techniques and aim to predict human annotations (explicit saliency judgment) and those attempting to predict fixations. We then compared both types of models over two types of data and concluded that fixation prediction models outperform salient object detection models for fixation prediction while being as good as those models for detecting the most salient object in a scene.

Acknowledgments

This work was supported by the National Science Foundation (Grant No. CMMI-1235539), the Army Research Office (W911NF-11-1-0046 and W911NF-12-1-0433), and US Army (W81XWH-10-2-0076). The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.

References

Achanta, R., Hemami, S., Estrada, F., & Süsstrunk, S. (2009). Frequency-tuned salient region detection. In *IEEE conference on computer vision and pattern recognition*.

Alpert, S., Galun, M., Basri, R., & Brandt, A. (2007). Image segmentation by probabilistic bottom-up aggregation and cue integration. In *IEEE conference on computer vision and pattern recognition*.

Ballard, D., Hayhoe, M., & Pelz, J. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience, 7*, 66–80.

Baluch, F., & Itti, L. (2011). Mechanisms of top-down attention. *Trends in Neurosciences, 34*, 210–224.

Berg, A., Berg, T., Daume III, H., Dodge, J., Goyal, A., Han, X., & et al. (2012). Understanding and predicting importance in images. In *IEEE conference on computer vision and pattern recognition*.

Borji, A., Sihite, D. N., & Itti, L. (2011). Computational modeling of top-down visual attention in interactive environments. In *Proceedings of the British machine vision conference (BMVC)* (pp. 85.1–85.12).

Borji, A., Sihite, D. N., & Itti, L. (2012). Salient object detection: A benchmark. In *European conference on computer vision (ECCV)*.

Borji, A., & Itti, L. (2013). State-of-the-art in modeling visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.

Borji, A., Sihite, D. N., & Itti, L. (2013a). Objects do not predict fixations better than early saliency; reanalysis of Einhäuser et al.'s data. *Journal of Vision*.

Borji, A., Sihite, D. N., & Itti, L. (2013b). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*.

Bruce, N., & Tsotsos, J. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision, 9*, 1–24.

Buswell, G. (1935). *How people look at pictures*. Chicago: University of Chicago Press.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research, 51*, 1484–1525.

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision, 9*.

Chang, K., Liu, T., Chen, H., & Lai, S. (2011). Fusing generic objectness and visual saliency for salient object detection. In: *International conference on computer vision (ICCV)*.

Cheng, M., Zhang, G., Mitra, N., Huang, X., & Hu, S. (2011). Global contrast based salient region detection. In *IEEE conference on computer vision and pattern recognition*.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*, 193–222.

Duncan, J. (1984). Selective attention and the organization of visual information. *ournal of Experimental Psychology: General, 113*, 501–517.

Ehinger, K., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modeling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition, 17*, 945–978.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision, 8*, 1–26.

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision, 8*, 3.1–15.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision, 8*, 1–17.

Garcia-Diaz, A., Fdez-Vidal, X. R., Pardo, X. M., & Dosil, R. (2012). Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing, 30*, 51–64.

Goferman, S., Zelnik-Manor, L., & Tal, A. (2010). Context-aware saliency detection. In *IEEE conference on computer vision and pattern recognition*.

Green, D., & Swets, J. (1966). *Signal detection theory and psychophysics*. New York: John Wiley.

Guo, C., & Zhang, L. (2010). A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Transactions on Image Processing, 9*.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. *Advances in Neural Information Processing Systems (NIPS), 19*, 545–552.

Hayhoe, M. (2000). Vision using routines: A functional account of vision. *Visual Cognition, 7*, 43–64.

Henderson, J. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences, 7*, 498–504.

Henderson, J. M, Brockmole, J. R, Castelhano, M. S, & Mack, M. (2007). Visual saliency does not account for eye movements during search in real-world scenes. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 537–562). Oxford: Elsevier.

Henderson, J., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50*, 243–271.

Henderson, J., Malcolm, G., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin and Review, 16*, 850–856.

Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Hou, X., & Zhang, L. (2008). Dynamic visual attention: Searching for coding length increments. *Advances in Neural Information Processing Systems (NIPS)*, 681–688.

Hwang, A., Wang, H., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision Research, 51*, 1192–1205.

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 145–152).

Itti, L. (2004). Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing, 13*, 1304–1318.

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition, 12*, 1093–1123.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*, 1489–1506.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2*, 194–203.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*, 1254–1259.

Jiang, H., Wang, J., Yuan, Z., Liu, T., Zheng, N., & Li, S. (2011). Automatic salient object segmentation based on context and shape prior. In *British machine vision conference (BMVC)*.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *International conference on computer vision (ICCV)*.

Kanan, C., & Cottrell, G. (2010). Robust classification of objects, faces, and flowers using national image. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4*, 219–227.

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research, 41*, 3559–3565.

Land, M. F., & Lee, D. N. (1994). Where we look when we steer. *Nature, 369*, 742–744.

Liu, T., Sun, J., Zheng, N., Tang, X., & Shum, H. (2007). Learning to detect a salient object. *IEEE Conference on Computer Vision and Pattern Recognition*.

Mack, S., & Eckstein, M. (2011). Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. *Journal of Vision, 11*.

Mackworth, N., & Morandi, A. (1967). The gaze selects informative details within pictures. *Perception & Psychophysics, 2*, 547–552.

Mannan, S. K., Kennard, C., & Husain, M. (2009). The role of visual salience in directing eye movements in visual object agnosia. *Current Biology, 19*.

Marchesotti, L., Cifarelli, C., & Csurka, G. (2009). A framework for visual saliency detection with applications to image thumbnailing. In *International conference on computer vision (ICCV)*, 2232–2239.

Masciocchi, C., Mihalas, S., Parkhurst, D., & Niebur, E. (2009). Everyone knows what is interesting: Salient locations which should be fixated. *Journal of Vision, 9*, 1–22.
Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature, 434*, 387–391.
Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research, 45*, 205–231.
Navalpakkam, V., & Itti, L. (2007). Search goal tunes visual features optimally. *Neuron, 53*, 605–617.
Navalpakkam, V., Koch, C., Rangel, A., & Perona, P. (2010). Optimal reward harvesting in complex perceptual environments. *Proceedings of the National Academy of Sciences, 107*, 5232–5237.
Nothdurft, H. (2005). Salience of feature contrast. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of attention* (pp. 233–239). Burlington, MA: Elsevier.
Nuthman, A., & Henderson, J. (2010). Object-based attentional selection in scene viewing. *Journal of Vision, 10*.
Pajak, M., & Nuthmann, A. (2013). Object-based saccadic selection during scene perception: Evidence from viewing position effects. *Journal of Vision, 13*.
Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*, 107–123.
Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research, 45*, 2397–2416.
Pomplun, M. (2006). Saccadic selectivity in complex visual search displays. *Vision Research, 46*, 1886–1900.
Posner, M. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32*, 3–25.
Powers, A. S., Basso, M. A., & Evinger, C. (2013). Blinks slow memory-guided saccades. *Journal of Neurophysiology, 109*, 734–741.
Rajashekar, J., Bovik, L., & Cormack, A. (2006). Visual search in noise: Revealing the influence of structural cues by gaze-contingent classification image analysis. *Journal of Vision, 17*, 379–386.
Rayner, K. (1979). Eye guidance in reading: Fixation locations within words. *Perception, 8*, 21–30.
Reinagel, P., & Zador, A. (1999). Natural scenes at the center of gaze. *Network, 10*, 341–350.
Russell, B., Torralba, A., Murphy, K., & Freeman, W. (2008). Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision, 77*, 157–173.
Seo, H., & Milanfar, P. (2009). Static and space-time visual saliency detection by self-resemblance. *Journal of Vision, 9*, 1–27.
Shen, J., & Itti, L. (2012). Top-down influences on visual attention during listening are modulated by observer sex. *Vision Research, 65*, 62–76.
Spain, M., & Perona, P. (2010). Measuring and predicting object importance. *International Journal of Computer Vision (IJCV), 99*, 59–76.
Tatler, B. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*, 1–17.
Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45*, 643–659.
Tatler, B., Hayhoe, M., Land, M., & Ballard, D. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision, 11*, 1–23.
Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113*, 766–786.
Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12*, 97–136.
Triesch, J., Ballard, D., Hayhoe, M., & Sullivan, B. (2003). What you see is what you need. *Journal of Vision, 3*, 86–94.
Tseng, P., Carmi, R., Cameron, I., Munoz, D., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision, 9*, 1–16.
Walther, D., Rutishauser, U., Koch, C., & Perona, P. (2005). Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Computer Vision and Image Understanding, 100*, 41–63.
Wischnewski, M., Belardinelli, A., & Schneider, W. (2010). Where to look next? Combining static and dynamic proto-objects in a TVA-based model of visual attention. *Cognitive Computation*, 326–343.
Wolfe, J. (1998). Visual search. In H. Pashler (Ed.), *Attention* (pp. 13–71). Hove, UK: Psychology Press.
Wolfe, J., & Horowitz, T. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience, 5*, 1–7.
Wright, R., & Ward, L. (2008). *Orienting of attention*. Oxford University Press.
Yarbus, A. (1967). *Eye movements and vision*. New York: Plenum.
Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review, 115*, 787–835.
Zetzsche, C. (2005). Natural scene statistics and salient visual features. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of attention*. London: Elsevier.
Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision, 8*.